REMARKS ON DRIFT ESTIMATION FOR DIFFUSION PROCESSES

YVO POKERN*, ANDREW M. STUART†, AND ERIC VANDEN-EIJNDEN‡

Abstract. In applications such as molecular dynamics it is of interest to fit Smoluchowski and Langevin equations to data. Practitioners often achieve this by a variety of seemingly *ad hoc* procedures such as fitting to the empirical measure generated by the data, and fitting to properties of auto-correlation functions. Statisticians, on the other hand, often use estimation procedures which fit diffusion processes to data by applying the maximum likelihood principle to the path-space density of the desired model equations, and through knowledge of the properties of quadratic variation. In this note we show that these procedures used by practitioners and statisticians to fit drift functions are, in fact, closely related and can be thought of as two alternative ways to regularize the (singular) likelihood function for the drift. We also present the results of numerical experiments which probe the relative efficacy of the two approaches to model identification and compare them with other methods such as the minimum distance estimator.

Key words. parameter estimation, diffusion process, nonparametric estimation, maximum likelihood principle, minimum distance estimator, reversible diffusion process, molecular dynamics, Langevin equation.

AMS subject classifications. 62M05 Markov processes: estimation 65C30 Stochastic differential and integral equations

1. Introduction. In many applications (such as molecular dynamics, econometrics, atmospheric sciences and signal processing) it is of interest to fit a diffusion process to a time-series. The data may come from experiments, or from the numerical simulation of larger and more complex models, either deterministic or stochastic. The objective of the present paper is to discuss some issues that arise when applying a maximum likelihood inference method to this problem. In so doing, we will highlight some connections between this approach, favored by statisticians, and other approaches used in the physics and chemistry literature.

To introduce the maximum likelihood inference framework and some of the issues that we will discuss, it is useful to consider first the specific case when it is known that the data is consistent with an Itô stochastic differential equation of the form:

$$\dot{X}_t = -\nabla V_0(X_t) + \sqrt{2\beta^{-1}}\,\dot{W}_t \tag{1.1}$$

This equation is often referred to as the Smoluchowski or overdamped Langevin equation in the chemical-physics literature. Precise statements of the observations about this problem given here will be provided in Section 2.1. In Section 2.2 we consider general reversible diffusions and in Section 3 the (non-reversible) second order Langevin equation.

In equation (1.1), W_t is a standard d-dimensional Brownian motion in $\mathbb{R}^d$, $\beta > 0$ is a constant playing the role of the inverse temperature, and $V_0 : \mathbb{R}^d \to \mathbb{R}$ is a potential which we assume C^2, bounded from below and with a growth condition at infinity to guarantee that $e^{-\beta V_0}$ is integrable. In this case, the process defined by (1.1) is ergodic with respect to the Boltzmann-Gibbs measure associated with V_0 whose density is

$$\rho_0(x) = Z^{-1} e^{-\beta V_0(x)} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} e^{-\beta V_0(x)} dx. \tag{1.2}$$

* Department of Statistics, University of Warwick, Coventry CV4 7AL, England.
† Mathematics Institute, University of Warwick, Coventry CV4 7AL, England.
‡ Courant Institute, New York University, New York, USA.

We assume that β is known and that we wish to estimate the potential V from the data, i.e. from a sample path $\{X_t\}_{t\in[0,T]}$ for some $T > 0$. For the time being we assume that a continuous sample of the path is available; later on in the paper, we will also discuss the problem when X_t is sampled at discrete times. To see how the problem of estimating V given β can be cast into a maximum likelihood inference problem, let Z_t solve (1.1) for $V_0 \equiv 0$ so that

$$\dot{Z}_t = \sqrt{2\beta^{-1}}\,\dot{W}_t, \tag{1.3}$$

and let $\mathbb{P}$ and $\mathbb{Q}$ be the path-space measures generated on $[0,T]$ by (1.1) and (1.3) respectively. Then these measures are absolutely continuous with Radon-Nikodym derivative

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp(-TI_T(X)) \tag{1.4}$$

where

$$I_T(X) = \frac{\beta}{4T}\int_0^T \left(|\nabla V_0(X_t)|^2 dt + 2\langle \nabla V_0(X_t), dX_t\rangle\right), \tag{1.5}$$

and $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product on $\mathbb{R}^d$ and $|\cdot|$ the Euclidean norm, and the integral with respect to dX_t is to be understood in the Itô sense. The functional $I_T(X)$ given by equation (1.5) is proportional to the negative logarithm of the probability density of the path $\{X_t\}_{t\in[0,T]}$ with respect to the measure on path-space generated by (1.3). When a single path $\{X_t\}_{t\in[0,T]}$ is given, if we evaluate (1.5) with potential V rather than V_0, this object becomes a functional of V. This functional is the negative of the log likelihood function for V:

$$\mathcal{I}_T(V) = \frac{\beta}{4T}\int_0^T \left(|\nabla V(X_t)|^2 dt + 2\langle \nabla V(X_t), dX_t\rangle\right) \tag{1.6}$$

Thus, it is natural to try to minimize (1.6) over V to obtain the maximum likelihood estimator (MLE) for this function. Indeed, using (1.1), (1.6) can be written as

$$\begin{aligned}\mathcal{I}_T(V) = &\frac{\beta}{4T}\int_0^T \left(|\nabla V(X_t)|^2 - 2\langle \nabla V(X_t), \nabla V_0(X_t)\rangle\right) dt \\ &+ \frac{\sqrt{2\beta}}{2T}\int_0^T \langle \nabla V(X_t), dW_t\rangle\end{aligned} \tag{1.7}$$

Letting $T \to \infty$, the stochastic integral in this expression tends to 0 almost surely (a.s.), whereas the time integral converges a.s. toward an expectation with respect to the equilibrium measure with density (1.2). In other words, as $T \to \infty$, $\mathcal{I}_T(V)$ converges a.s. to the functional $\mathcal{I}_\infty(V)$ given by

$$\mathcal{I}_\infty(V) = \frac{\beta}{4}\int_{\mathbb{R}^d} \left(|\nabla V(x)|^2 - 2\langle \nabla V(x), \nabla V_0(x)\rangle\right) \rho_0(x)dx \tag{1.8}$$

This functional is quadratic and convex in ∇V and, by completing the square, it is clearly minimized when $\nabla V = \nabla V_0$, i.e. when $V = V_0 + C$ where C is an arbitrary (and irrelevant) constant. Thus the MLE for $-\nabla V_0$ given by maximizing the limiting functional (1.8) is indeed the actual drift in (1.1).

The problem, however, is that the the data $\{X_t\}_{t\in[0,T]}$ is finite, $T<\infty$, i.e. we are obliged to work with (1.6) and have no access to its infinite time limit (1.8). To see what problems this creates, let us first put (1.6) in a more convenient form by converting the Itô stochastic integral $\langle \nabla V(X_t), dX_t\rangle$ into the Stratonovich integral using

$$\langle \nabla V(X_t), \circ dX_t\rangle = \langle \nabla V(X_t), dX_t\rangle + \beta^{-1}\Delta V(X_t)dt.$$

Since $\langle \nabla V(X_t), \circ dX_t\rangle = dV(X_t)$ this gives

$$\mathcal{I}_T(V) = \frac{\beta}{2T}\big(V(X_T)-V(X_0)\big) + \frac{\beta}{4T}\int_0^T \big(|\nabla V(X_t)|^2 - 2\beta^{-1}\Delta V(X_t)\big)\,dt \qquad (1.9)$$

The time integral in (1.9) can be transformed into a configuration integral using the occupation measure μ_T defined such that, for any Borel set $B\subset\mathbb{R}^d$, one has

$$\mu_T(B) = \frac{1}{T}\int_0^T \mathbf{1}_B(X_t)dt \qquad (1.10)$$

where $\mathbf{1}_B(x)$ is the indicator function of the set B. The measure μ_T is the finite time equivalent of the equilibrium measure $\rho_0(x)dx$ entering (1.8). Using μ_T, we can write (1.9) as

$$\mathcal{I}_T(V) = \frac{\beta}{2T}\big(V(X_T)-V(X_0)\big) + \frac{\beta}{4}\int_{\mathbb{R}^d} \big(|\nabla V(x)|^2 - 2\beta^{-1}\Delta V(x)\big)\,\mu_T(dx). \qquad (1.11)$$

This expression (1.11) makes it apparent why an attempt to directly minimize this functional over V is a bad idea. When $d=1$, the occupation measure μ_T has the scaled local time L_T^x/T of the process $\{X_t\}_{t\in[0,T]}$ as a density, but L_T^x is only Hölder continuous up to $C^{0,\frac{1}{2}}(\mathbb{R})$. Indeed, L_T^x has the fine-scale properties of a diffusion process (cf. the Ray-Knight description of Brownian local times). In the appendix, we show that (1.11) evaluated with $\mu_T(dx)=w(x)dx$ where $w(x)$ is a one dimensional Brownian motion, is not bounded from below. When $d>1$, μ_T is singular with respect to the Lesbegue measure since it is supported on $\{X_t\}_{t\in[0,T]}$. Thus $\mathcal{I}_T(V)$ must be regularized in some way to become useful. There are at least three obvious ways to perform such a regularization.

1. The first way, which we will not discuss in this paper, is to adopt a Bayesian non-parametric approach in which a prior measure on V is introduced that is supported on sufficiently regular functions only. By sampling from this measure and using the exponential of the negative of (1.6) or, equivalently, (1.11) as reweighting density, it is possible to sample the posterior distribution of V given the data $\{X_t\}_{t\in[0,T]}$. This approach is discussed in [16] and we refer the reader to that paper for details.

2. A second way to regularize (1.11) is to assume a parametric form for V, e.g. as a linear combination of smooth basis functions $f_i(x)$,

$$V(x,\theta) = \sum_{j=1}^N \theta_i f_i(x). \qquad (1.12)$$

where $\theta_1,\dots,\theta_N$ are weights. By substituting (1.12) into (1.6), one is left with a quadratic function of $\theta=(\theta_1,\dots,\theta_N)\in\mathbb{R}^N$

$$\mathcal{I}_T(\theta) = \frac{\beta}{4T}\int_0^T \left(\Big|\sum_{i=1}^N \theta_i \nabla f_i(X_t)\Big|^2 dt + 2\sum_{i=1}^N \theta_i \langle \nabla f_i(X_t), dX_t\rangle\right). \qquad (1.13)$$

For appropriate choice of $f_i(x)$, this quadratic function of θ is convex and therefore has a unique minimum $\hat{\theta}$ which can be found by solving a linear algebraic system. This approach is the one often adopted in the statistics literature to identify a parametric approximation to the MLE of V. We will refer to it as the **parametric approach**. Notice that it is crucial for this approach to work that the sum in (1.12) be finite, since it is this which regularizes the functional (1.6); the actual (non-parametric) MLE for V will not exist in general.

3. A third way to regularize (1.11) is to regularize the measure $\mu_T(dx)$ and replace it by $\rho_T(x)dx$, where $\rho_T(x) > 0$ is a smooth probability density function. With this substitution, (1.11) becomes

$$\mathcal{I}_T(V) = \frac{\beta}{2T}\big(V(X_T) - V(X_0)\big) + \tilde{\mathcal{I}}_T(V) \tag{1.14}$$

where

$$\tilde{\mathcal{I}}_T(V) = \frac{\beta}{4}\int_{\mathbb{R}^d} \left(|\nabla V(x)|^2 - 2\beta^{-1}\Delta V(x)\right)\rho_T(x)dx \tag{1.15}$$

If T is large enough, it is reasonable to neglect the first term on the right-hand side of (1.14), i.e. approximate $\mathcal{I}_T(V)$ by $\tilde{\mathcal{I}}_T(V)$. To identify the minimizer of $\tilde{\mathcal{I}}_T(V)$, note that if $\rho_T(x) > 0$ and for potentials V such that $\lim_{x\to\infty} \nabla V(x)\rho_T(x) = 0$, an integration by parts yields

$$\begin{aligned}\tilde{\mathcal{I}}_T(V) &= \frac{\beta}{4}\int_{\mathbb{R}^d} \left(|\nabla V(x)|^2 + 2\beta^{-1}\langle \nabla V(x), \nabla \log \rho_T(x)\rangle\right)\rho_T(x)dx \\ &= \frac{\beta}{4}\int_{\mathbb{R}^d} \left(|\nabla V(x) + \beta^{-1}\nabla \log \rho_T(x)|^2 - \beta^{-2}|\nabla \log \rho_T(x)|^2\right)\rho_T(x)dx.\end{aligned} \tag{1.16}$$

This last expression shows that the minimizer of $\tilde{\mathcal{I}}_T(V)$ is unique up to a constant and given by

$$\hat{V}(x) = -\beta^{-1}\log \rho_T(x) + C' \tag{1.17}$$

where C' is an arbitrary constant. This expression for V is the one usually adopted in the physics and chemistry literature and we will refer to it as the **non-parametric approach** since (1.16) and, hence, (1.17) involve no direct parametrization of V. Notice however that this approach leaves as an auxiliary problem the issue of determining $\rho_T(x)$. Thus, rather than removing the issue of parameterisation, it merely displaces it to $\rho_T(x)$. This density can itself be obtained by minimization of some appropriate functional (see (4.10) in the section on Numerics).

The calculations above show some of the issues that arise when a maximum likelihood inference method is applied to estimate the drift (here $-\nabla V_0(X_t)$) in a diffusion (here (1.1)). They also uncover a connection between the maximum likelihood inference method often adopted by statisticians and the procedure of fitting V to some empirical equilibrium density which is used by chemists and physicists. In the remainder of this paper we will generalize this connection. Specifically:

1. In Section 2, we will clean up the calculations above and prove the facts that we just listed. We will also outline how these calculations could be generalised to a generic time-reversible diffusion and indicate that a connection between the maximum

likelihood inference and the procedure of fitting the drift to some empirical equilibrium density may exist in this case as well.

2. In section 3, we will generalize these conclusions to a specific non-reversible diffusion of great practical importance, namely the Langevin equation, a hypo-elliptic diffusion process found by coupling a Hamiltonian system to a heat bath via white noise and damping.

3. In section 4, we will perform a series of numerical experiments to illustrate our results and discuss a series of remaining issues: What is the influence of neglecting the boundary terms in (1.14)? What happens when the data is sampled at discrete times (in this case (1.6) and (1.9), and hence (1.6) and (1.11) are no longer equivalent)? What are the options to estimate $\rho_T(x)$ in (1.14)?

2. Drift inference for time-reversible processes.

2.1. Smoluchowski equation. In this section we make precise the results in the Introduction.

First we analyze some properties of the log likelihood function $\mathcal{I}_T(V)$, written either as in (1.6) or (1.11). We start by stating a theorem which indicates that attempting to minimize (1.11) directly may be ill-advised. We do this in the special case $d = 1$ and where the domain of integration is restricted to $[0, 1]$ and boundary terms are neglected, i.e. we consider the funtional:

$$\mathcal{I}_B(b) = \int_0^1 \left(b^2(x) - b'(x)\right) \mu(dx) \tag{2.1}$$

for $b \in H^1(0, 1)$.

THEOREM 2.1. *If $\mu(dx)$ in (2.1) is absolutely continuous with respect to Lebesgue measure with density given by a realisation of the Brownian bridge, then the functional $\mathcal{I}_B(b)$ is almost surely not bounded below for $b \in H^1(0, 1)$.*

Proof. See the appendix. □

While singular in the sense above when $T < \infty$, the log likelihood function $\mathcal{I}_T(V)$ has a nice limit as $T \to \infty$, as shown by the following:

THEOREM 2.2. *Assume that there exist $C_1, C_2 > 0$ such that for all $x \in \mathbb{R}^d$*

$$C_1 + \langle x, \nabla V_0(x)\rangle \geq C_2|x|^2$$

and that both $V_0(x)$ and $V(x)$ are polynomially bounded. Then as $T \to \infty$, the functional $\mathcal{I}_T(V)$ in (1.6) converges a.s. to the functional $\mathcal{I}_\infty(V)$ defined in (1.8).

This theorem is a consequence of the following lemma:

LEMMA 2.3. *Under the assumptions of Theorem 2.2, equation (1.1) is ergodic with respect to the equilibrium measure with the density (1.2) and*

$$\limsup_{t\to\infty} \frac{\sqrt{2\beta}}{2T} \int_0^T \langle \nabla V(X_t), dW_t\rangle = 0 \quad a.s. \tag{2.2}$$

Proof. The ergodicity follows from [14]. Theorem 5.5 of Chapter 2 in [13] implies that

$$\limsup_{t\to\infty} \frac{|X_t|}{\sqrt{\log t}} \leq \sqrt{\frac{2e}{C_2\beta}} \quad \text{a.s.} \tag{2.3}$$

Let $\mathcal{L}$ denote the generator of the process (1.1). By the Itô formula we have

$$V(X_t) - V(X_0) = \int_0^T (\mathcal{L}V)(X_t)dt + \int_0^T \langle \nabla V(X_t), dW_t \rangle \quad \text{a.s.} \tag{2.4}$$

Now divide (2.4) by T.

$$\frac{1}{T}\left(V(X_t) - V(X_0)\right) = \frac{1}{T}\int_0^T (\mathcal{L}V)(X_t)dt + \frac{1}{T}\int_0^T \langle \nabla V(X_t), dW_t \rangle.$$

The bound (2.3) shows that the term $\frac{1}{T}\left(V(X_t) - V(X_0)\right)$ tends to zero. Also, by ergodicity

$$\frac{1}{T}\int_0^T (\mathcal{L}V)(X_t)dt \to \int_{\mathbb{R}^d} \mathcal{L}V(x)\rho_0(x)dx = 0$$

since $\mathcal{L}^*\rho_0 = 0$. Thus, (2.2) follows. □

Next we analyze the parametric log likelihood function (1.13) used in the **parametric approach**. We have

THEOREM 2.4. *Let $F = \{f_{ij}\}$ be the matrix with entries*

$$f_{ij} = \frac{1}{T}\int_0^T \langle \nabla f_i(X_t), \nabla f_j(X_t) \rangle dt, \qquad i,j = 1,\dots,N \tag{2.5}$$

and assume that F is positive definite. Then (1.13) has a unique minimizer. In addition, this minimizer is then given by

$$\hat{\theta} = F^{-1}h \tag{2.6}$$

where h is the vector with components

$$h_i = -\frac{1}{T}\int_0^T \langle \nabla f_i(X_t), dX_t \rangle dt, \qquad i = 1,\dots,N. \tag{2.7}$$

Furthermore, if the ∇f_i are polynomially bounded then $\lim_{T\to\infty} F$ exists and is almost surely invertible.

Proof. Immediate. □

Finally, we analyze the properties of the approximate log likelihood function $\tilde{\mathcal{I}}_T(V)$ in (1.14) used in the **non-parametric approach**. An immediate consequence of (2.3) is that the boundary term in (1.14) is negligible.

THEOREM 2.5. *Under the assumptions of Theorem 2.2, we have, for any $\varepsilon > 0$, that*

$$\limsup_{t\to\infty} \frac{V(X_t)}{t^\varepsilon} = 0 \quad \textit{a.s.} \tag{2.8}$$

The next theorem shows that the minimization problem associated with (1.15) has a unique solution as long as the density ρ_T in this functional satisfies some requirements. To be able to state it more neatly, we introduce the space $\mathcal{V}$ as follows. For any open and bounded subset $U \subset \mathbb{R}^d$ define

$$\mathcal{V}(U) = \left\{ V \in H^1(U) : \int_U V(x)dx = 0 \right\}$$

THEOREM 2.6. *Let $\rho_T : \mathbb{R}^d \to \mathbb{R}$ be smooth, $\rho_T \in C^\infty(\mathbb{R}^d)$. Furthermore, let U be a bounded open subset of $\mathbb{R}^d$ and let ρ_T be bounded below on U: $\exists \varepsilon > 0$, $\forall x \in U$: $\rho_T(x) > \varepsilon$. Then the minimizer of*

$$\inf_{V \in \mathcal{V}(U)} \frac{\beta}{4} \int_U \Big(|\nabla V(x)|^2 - 2\beta^{-1} \Delta V(x) \Big) \rho_T(x) dx \tag{2.9}$$

is unique and given by

$$\hat{V}(x) = -\beta^{-1} \log \rho_T(x) + C \quad \textit{where} \quad C = \beta^{-1} \int_U \rho(x) dx \tag{2.10}$$

The theorem can be proved using results from [4] but the proof can also be carried out by directly completing the square. The basic idea was given in the developments made in (1.16).

2.2. The generic time-reversible diffusion process. In this section, we assume that the data $\{X_t\}_{t \in [0,T]}$ has been generated by the following Itô stochastic differential equation:

$$\dot{X}_t = b_0(X_t) + \sigma_0(X_t) dW_t \tag{2.11}$$

where $b_0 : \mathbb{R}^d \to \mathbb{R}^d$ is the drift coefficient, $\sigma_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the diffusion coefficient, and W_t is a standard d-dimensional Brownian motion. We assume that the diffusion coefficient $\sigma_0(x)$ is known and satisfies

$$\exists C > 0 \ : \ \langle \eta, \sigma_0 \sigma_0^T(x) \eta \rangle \geq C |\eta|^2 \quad \forall x, \eta \in \mathbb{R}^d \tag{2.12}$$

and that we wish to estimate the drift $b_0(x)$. We also assume that the process generated by (2.11) is ergodic with respect to the equilibrium measure with density $\rho_0(x)$ (which we do not know *a priori*) and that this process is time-reversible. This last assumption means that

$$\begin{array}{l} \{X_{t-T/2}\}_{t \in [-T/2, T/2]} \text{ and } \{X_{T/2-t}\}_{t \in [-T/2, T/2]} \text{ are equivalent in law} \\ \text{in the limit as } T \to \infty. \end{array} \tag{2.13}$$

The time-reversibility also implies that $b_0(x)$, $a_0(x) = \sigma_0 \sigma_0^T(x)$ and $\rho_0(x)$ are related as

$$0 = b_0 \rho_0 - \tfrac{1}{2} \mathrm{div}(a_0 \rho_0) \tag{2.14}$$

which expresses that a time-reversible process has no probability current at equilibrium. Note that since ρ_0 is unknown to us (only σ_0 and hence $a_0 = \sigma_0 \sigma_0^T$ are assumed to be available), (2.14) cannot be used *a priori* to determine b_0. Nevertheless, the **non-parametric approach** would be to simply approximate ρ_0 in (2.14) by some empirical density ρ_T and thereby obtain an estimate for b. Next we show that this approach is closely related to the **parametric approach** in that both approaches correspond to minimizing a different regularization of the likelihood functional for b_0.

Proceeding as in the Introduction, we can derive the negative of the log likelihood functional for the unknown drift b given the data $\{X_t\}_{t \in [0.t]}$. Up to an irrelevant constant, this functional is

$$\mathcal{I}_T(b) = \frac{1}{T} \int_0^T \Big(|b(X_t)|^2_{a_0(X_t)} dt - 2 \langle b(X_t), dX_t \rangle_{a_0(X_t)} \Big) \tag{2.15}$$

where we introduced the following inner product and norm on the tangent space at $x \in \mathbb{R}^d$:

$$\begin{aligned} \langle \eta, \xi \rangle_{a_0(x)} &= \langle \eta, {a_0}^{-1}(x)\xi \rangle \qquad \forall \eta, \xi \in \mathbb{R}^d, \\ |\eta|^2_{a_0(x)} &= \langle \eta, \eta \rangle_{a_0(x)} \qquad\quad \forall \eta \in \mathbb{R}^d. \end{aligned} \tag{2.16}$$

This inner product and the norm are well defined since $a_0(x)$ is invertible at every $x \in \mathbb{R}^d$ by assumption (2.12).

As in (1.6), the log likelihood function (2.15) for b is unbounded below in general if the data is finite, $T < \infty$. We can however proceed as for the Smoluchowski equation (1.1) along the following lines:

1. If we let $T \to \infty$, (2.15) tends to a functional whose unique minimizer is b_0.

2. If we parametrize b by the following form suggested by (2.14)

$$b(x) = \tfrac{1}{2}\mathrm{div}\, a_0(x) - \tfrac{1}{2}a_0(x)\nabla V(x, \theta) \tag{2.17}$$

with $V(x, \theta)$ as in (1.12) (thus $V(x, \theta)$ is approximating $-\log \rho_0$), (2.15) becomes a quadratic and convex function for $\theta = (\theta_1, \ldots, \theta_N)$ whose unique minimizer can be determined by solving a linear algebraic problem. This is the **parametric approach**.

3. There is an alternative way to regularize (2.15) which involves transforming the time integral in (2.15) into an expectation with respect to the occupation measure (1.10), and approximating $\mu_T(dx)$ by $\rho_T(x)dx$ where $\rho_T(x)$ is some smooth density. Then the minimizer of this regularized log likelihood function is unique and related to ρ_T in the same way as b_0 is related to ρ_0 in (2.14). This is the **non-parametric approach**.

Let us analyze in more detail the statements made in these three points. The statement made in point 1 is a simple consequence of using (2.11) to re-write (2.15) as

$$\begin{aligned} \mathcal{I}_T(b) = \frac{1}{T}\int_0^T \Big(|b(X_t)|^2_{a_0(X_t)} - 2\langle b(X_t), b_0(X_t)\rangle_{a_0(X_t)} \Big)\, dt \\ - \frac{2}{T}\int_0^T \langle b(X_t), \sigma_0(X_t) dW_t \rangle_{a_0(X_t)} \end{aligned} \tag{2.18}$$

In the limit as $T \to \infty$, we would expect, by exploiting time-reversibility, that the stochastic integral converges a.s. to zero; this is exactly what happens for the Smoluchowski equation (see Lemma 2.3). By ergodicity, the first integral converges a.s. towards an expectation with respect to the equilibrium distribution with density ρ_0. Thus, $\mathcal{I}_T(b)$ is expected to converge almost surely towards the functional $\mathcal{I}_\infty(b)$ given by

$$\mathcal{I}_\infty(b) = \int_{\mathbb{R}^d} \Big(|b(x)|^2_{a_0(x)} - 2\langle b(x), b_0(x)\rangle_{a_0(x)} \Big)\, \rho_0(x) dx \tag{2.19}$$

If $\rho_0(x) > 0$, completing the square shows that the minimizer of this functional is unique and given by $b(x) = b_0(x)$, as needed. Of course, (2.19) is unavailable in practice since the data is finite.

Consider now the statement made in point 2. If we insert (2.17) into (2.15) and neglect all the irrelevant terms independent of θ, as well as an overall multiplicative constant, we arrive at

$$\mathcal{I}_T(\theta) = \theta^T \bar{F} \theta - 2\theta^T \bar{h} \tag{2.20}$$

where $\bar{F} = \{\bar{f}_{ij}\}$ is the matrix with entries

$$\bar{f}_{ij} = \frac{1}{T}\int_0^T \langle \nabla f_i(X_t), a_0(X_t)\nabla f_j(X_t)\rangle dt, \qquad i,j = 1,\dots,N \tag{2.21}$$

and $\bar{h}$ is the vector with components

$$\bar{h}_i = \frac{1}{T}\int_0^T \left(\langle \nabla f_i(X_t), \operatorname{div} a_0(X_t)\rangle dt - 2\langle \nabla f_i(X_t), dX_t\rangle\right), \quad i = 1,\dots,N \tag{2.22}$$

This is a quadratic function in θ which is strictly convex *iff* the matrix F is positive definite. If this is the case, (2.20) has a unique minimizer given by

$$\theta = \bar{F}^{-1}\bar{h} \tag{2.23}$$

These results are the equivalent for (2.11) of Theorem 2.4 for the Smoluchowski equation (1.1). Note that these results remain true even if the process defined by (2.11) is not time-reversible, since (2.20) remains the parametric approximation via (2.17) of the negative log likelihood function for b irrespective of whether the process is time-reversible or not.

To establish the statements made in point 3 above, we will use the following relation between the Itô integral in (2.15) and the corresponding Stratonovich integral

$$\begin{aligned}\int_0^T \langle b(X_t), dX_t\rangle_{a_0(X_t)} = &\int_0^T \langle b(X_t), \circ dX_t\rangle_{a_0(X_t)} \\ &+ \frac{1}{2}\int_0^T \left(\langle b(X_t), \operatorname{div} a_0(X_t)\rangle_{a_0(X_t)} - \operatorname{div} b(X_t)\right) dt\end{aligned} \tag{2.24}$$

Using this relation as well as the occupation measure μ_T of the process $\{X_t\}_{t\in[0,T]}$, (2.15) can be written at

$$\begin{aligned}\mathcal{I}_T(b) = &\int_{\mathbb{R}^d} \left(|b(x)|^2_{a_0(x)} + \operatorname{div} b(x) - \langle b(x), \operatorname{div} a_0(x)\rangle_{a_0(x)}\right)\mu_T(dx) \\ &- \frac{2}{T}\int_0^T \langle b(X_t), \circ dX_t\rangle_{a_0(X_t)}\end{aligned} \tag{2.25}$$

The stochastic integral in this expression is a correction term which we expect to vanish in the limit as $T \to \infty$, i.e. we would expect We have

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T \langle b(X_t), \circ dX_t\rangle_{a_0(X_t)} = 0 \quad \text{a.s.} \tag{2.26}$$

Thus, if we assume that T is large enough so that we can neglect the stochastic integral term in (2.25) and we approximate the occupation measure $\mu_T(x)$ by $\rho_T(x)dx$ where $\rho_T(x)$ is a smooth density with bounded support, we can approximate the log likelihood function (2.25) by

$$\tilde{\mathcal{I}}_T(b) = \int_{\mathbb{R}^d} \left(|b(x)|^2_{a_0(x)} + \operatorname{div} b(x) - \langle b(x), \operatorname{div} a_0(x)\rangle_{a_0(x)}\right)\rho_T(x)dx \tag{2.27}$$

This functional is the equivalent for (2.11) of the expression (1.15) for the Smoluchowski equation (1.1). Given the smoothness of the density we can perform the following partial integration:

$$\tilde{\mathcal{I}}_T(b) = \int_{\mathbb{R}^d} \Big(|b(x)|^2_{a_0(x)} \rho_T(x) - b(x) \cdot \nabla \rho_T(x) - \langle b(x), \mathrm{div} a_0(x) \rangle_{a_0(x)} \rho_T(x) \Big) \, dx, \tag{2.28}$$

where the boundary terms vanish since $\rho_T(\cdot)$ has bounded support. This functional has much nicer properties than the original $\mathcal{I}_T(b)$ in (2.15) as shown by the following result:

THEOREM 2.7. *Let U be a bounded open subset of $\mathbb{R}^d$ and assume that $\rho_T \in \mathcal{C}^\infty(U)$ is bounded below on U: $\exists \varepsilon > 0 : \rho_T(x) > \varepsilon \ \forall x \in U$. Furthermore, assume that $a_0(\cdot) \in \mathcal{C}^\infty(U)$ is positive definite symmetric everywhere on U and its lowest eigenvalue is bounded below: $\inf_{x \in U} \lambda_{\min}(a_0(x)) > 0$. Then for the functional*

$$\tilde{\tilde{\mathcal{I}}}_T(b) = \int_U \Big(|b(x)|^2_{a_0(x)} \rho_T(x) - b(x) \cdot \nabla \rho_T(x) - \langle b(x), \mathrm{div}\, a_0(x) \rangle_{a_0(x)} \rho_T(x) \Big) \, dx \tag{2.29}$$

the minimizer of

$$\inf_{b \in L^2(U)} \tilde{\tilde{\mathcal{I}}}_T(b)$$

is unique and given by

$$\tilde{b} = \tfrac{1}{2} \mathrm{div}(a_0 \rho_T)/\rho_T \qquad (x \in U) \tag{2.30}$$

Proof. Rewrite the functional $\tilde{\tilde{\mathcal{I}}}$ introducing an extra factor of ρ_T and a_0 in the middle term to recognise it as a quadratic form in b:

$$\tilde{\tilde{\mathcal{I}}}(b) = \int_U \left(|b(x)|^2_{a_0(x)} - \langle b(x), a_0(x) \frac{\nabla \rho_T(x)}{\rho_T(x)} \rangle_{a_0(x)} - \langle b(x), \mathrm{div} a_0(x) \rangle_{a_0(x)} \right) \rho_T(x) dx.$$

Now complete the square to obtain

$$\begin{aligned} \tilde{\tilde{\mathcal{I}}}(b) = \int_U \Big(& \left| b(x) - \frac{a_0(x)}{2} \frac{\nabla \rho_T(x)}{\rho_T(x)} - \frac{1}{2} \mathrm{div} a_0(x) \right|^2_{a_0(x)} \\ & - \left| \frac{a_0(x)}{2} \frac{\nabla \rho_T(x)}{\rho_T(x)} - \frac{1}{2} \mathrm{div} a_0(x) \right|^2_{a_0(x)} \Big) \rho_T(x) dx \end{aligned}$$

Since $\rho_T(\cdot)$ is strictly positive on U, this functional is minimised when

$$0 = b - \frac{a_0(x)}{2} \frac{\nabla \rho_T}{\rho_T} - \frac{1}{2} \mathrm{div} a_0. \qquad (x \in U) \tag{2.31}$$

This is an algebraic equation for b whose solution is (2.30). □

Relation (2.30) is the equivalent of (1.17) for a generic time-reversible process and shows how the **non-parametric approach** of deducing the drift coefficient from the equilibrium density and the diffusion coefficient can be generalized to this case.

An interesting consequence of the calculations above is that the time-ordering of the data is not very relevant for time-reversible processes. This is clear for the **non-parametric approach** based on (2.27) and leading to (2.30) in which only the empirical density $\rho_T(x)$ plays a role. Similarly, we expect that time-ordering plays only a small role in the **parametric approach** based on regularizing the maximum likelihood function leading to (2.20) via parametrization of the drift b. This conjecture will be verified in the numerical experiments of section 4.

3. Non-reversible processes: the Langevin equation. The calculations in section 2 rely heavily on the property that the process is time-reversible. In particular, for a non-reversible process, we would not expect 2.26 to hold in general, hence we will not be able to approximate the log likelihood function by (2.27) (the contribution from the stochastic integral term in (2.25) is missing). Another way to look at the problem is to realize that, for a non-reversible process, relation (2.14) is replaced by

$$j_0(x) = b_0\rho_0 - \tfrac{1}{2}\mathrm{div}(a_0\rho_0) \tag{3.1}$$

where $j_0(x)$ is a divergence-free vector field accounting for the non-zero equilibrium probability current of the non-reversible process. Equation (3.1) implies that it is not straightforward to generalize the **non-parametric approach** to non-reversible processes since, on top of the diffusion tensor a_0 and the equilibrium density ρ_0 (or some approximations thereof), we need an approximation of the current j_0 to deduce the drift b_0. This approximation of j_0 will not be available in general. Despite all this, in this section we show that the **non-parametric approach** can be generalized to a specific type of non-reversible processes which frequently arises in applications, and that this approach is again closely connected to the **parametric approach** for these processes. The specific type of non-reversible processes are those governed by the Langevin equation:

$$\ddot{Q}_t + \beta_0 D_0 \dot{Q}_t + \nabla V_0(Q_t) = \sqrt{2D_0}\dot{W}_t \tag{3.2}$$

where β_0 is the inverse temperature, D_0 is the diffusivity and W_t is a standard Brownian motion. (Thus the friction coefficient γ is related to β_0 and D_0 via the Einstein relation: $D_0 = \gamma/\beta_0$.) We assume that D_0 is known and we wish to find the potential V_0 and the inverse temperature β_0.

If we set $P_t = \dot{Q}_t$ (Q_t is referred to as position, P_t as momentum) then from (3.2) we obtain the following system of equations:

$$\begin{cases} \dot{Q}_t = P_t, \\ \dot{P}_t = -\beta_0 D_0 P_t - \nabla V_0(Q_t) + \sqrt{2D_0}\,\dot{W}_t. \end{cases} \tag{3.3}$$

Note that since the noise only enters the equation for P_t, (3.3) does not define an elliptic diffusion; it is, however, hypo-elliptic, see [14]. If one assumes that V_0 satisfies the assumptions in Theorem 2.2, the process generated by (3.3) is ergodic with respect to the equilibrium distribution with density

$$\varrho_0(q,p) = \rho_0(q)g_0(p) \tag{3.4}$$

where

$$\rho_0(q) = Z^{-1}e^{-\beta_0 V_0(q)}, \qquad g_0(q) = (2\pi\beta_0)^{-d/2}e^{-\frac{1}{2}\beta_0|p|^2} \tag{3.5}$$

Note in particular that the equilibrium distribution is Gaussian in the momentum coordinate.

The Radon-Nikodym derivative of the measure on path-space for (3.3) with respect to the measure generated by

$$\dot{P}_t = \sqrt{2D_0}\,\dot{W}_t. \tag{3.6}$$

is given by

$$\exp\Big(-\frac{T}{2D_0}I_T(Q,P)\Big) \tag{3.7}$$

Here

$$I_T(Q,P) = \frac{1}{2T}\int_0^T \big(|\beta_0 D_0 P_t + \nabla V_0(Q_t)|^2 dt + 2\langle \beta_0 D_0 P_t + \nabla V_0(Q_t), dP_t\rangle\big) \tag{3.8}$$

where it is understood that Q_t and P_t are related as $\dot{Q}_t = P_t$ as in (3.3). For fixed data $\{Q_t, P_t\}_{t\in[0,T]}$, we may evaluate (3.8) at V and β different from V_0 and β_0. The resulting functional is then the negative of the log likelihood function for V and β:

$$\mathcal{I}_T(V,\beta) = \frac{1}{2T}\int_0^T \big(|\beta D_0 P_t + \nabla V(Q_t)|^2 dt + 2\langle \beta D_0 P_t + \nabla V(Q_t), dP_t\rangle\big) \tag{3.9}$$

As in the Smoluchowski case, the log likelihood function (3.9) must be regularized to be useful. The simplest way is to parametrize $V(q)$ as in (1.12), in which case (3.9) reduces to a function of β and $\theta = (\theta_1, \ldots, \theta_N)$ which can then be minimized over these parameters. This is the **parametric approach**. Next we investigate another type of regularization of (3.9) leading to the equivalent of the **non-parametric approach**.

We begin by making a few transformations on (3.9). First, notice that an integration by parts using the Itô formula and $\dot{Q}_t = P_t$ shows that

$$\int_0^T \langle \nabla V(Q_t), dP_t\rangle = -\int_0^T \langle P_t, \nabla\nabla V(Q_t) P_t\rangle dt + \big[\langle \nabla V(Q_t), P_t\rangle\big]_0^T$$

$$\int_0^T \langle P_t, dP_t\rangle = \frac{1}{2}\big[|P_t|^2\big]_0^T - dD_0 T.$$

Thus

$$\begin{aligned}\mathcal{I}_T(V,\beta) = {} & \frac{1}{T}\Big[\beta D_0 |P_t|^2 + \langle \nabla V(Q_t), P_t\rangle\Big]_0^T \\ & - d{D_0}^2\beta + \frac{1}{2T}\int_0^T \big(|\beta D_0 P_t + \nabla V(Q_t)|^2 - 2\langle P_t, \nabla\nabla V(q) P_t\rangle\big)\, dt\end{aligned} \tag{3.10}$$

Under suitable conditions on the potentials V_0 and V, the boundary contributions from the two integrations by parts converge almost surely to zero as $T \to \infty$ as made precise in the following lemma:

LEMMA 3.1. *Assume that $\exists C_i > 0$ $i = 1, \ldots, 5$, where $C_1 < 1$ and $m \in \mathbb{Z}^+$ such that:*

- $\frac{1}{2}\langle \nabla V_0(q), q\rangle \geq C_1 V_0(q) + C_2|q|^2 - C_3 \quad \forall q \in \mathbb{R}^d$;
- $0 \leq |\nabla V(q)| \leq C_4\left[1 + |q|^{2m-1}\right] \quad \forall q \in \mathbb{R}^d$
- $0 \leq |\nabla V_0(q)| \leq C_5\left[1 + |q|^{2m-1}\right] \quad \forall q \in \mathbb{R}^d$.

Then there is a $C > 0$ such that:

$$\limsup_{t\to\infty} \frac{|P_t|^2 + |Q_t|^2}{\log t} \leq C \quad a.s.$$

and

$$\limsup_{t\to\infty} \frac{|\langle \nabla V(Q_t), P_t\rangle| + |P_t|^2}{t} = 0 \quad a.s.$$

Proof. Let $H(q,p)$ denote the following perturbed Hamiltonian:

$$H(q,p) = \frac{1}{2}|p|^2 + V(q) + D_0\beta_0\langle p, q\rangle + D_0^2\beta_0^2|q|^2 + 1.$$

Then

$$H(q,p) \geq 1 + \frac{1}{8}|p|^2 + \frac{D_0^2\beta_0^2}{3}|q|^2.$$

The arguments in Section 3 of [14] show that there exist $\xi_6, \xi_7, \xi_8, \xi_9 > 0$ such that:

$$\mathcal{L}H \leq \xi_6 - \xi_7 H$$

and

$$\left|\left\langle \nabla H, \begin{pmatrix} 0 \\ \sqrt{D_0} \end{pmatrix} \right\rangle\right|^2 \leq \xi_8 \left[|p| + |q|\right]^2 \leq \xi_9 H(q,p).$$

Thus, applying the Itô formula to $e^{\xi_7 t}H(q(t),p(t))$ and use of arguments similar to those in Theorem 5.5 of Chapter 2 in [13], but applied to $H(q,p)$ instead of $|p|^2+|q|^2$, give the first result. The second result follows since $\nabla V(q)$ is assumed polynomially bounded. □

The ergodicity of the process together with the lemma imply that as $T \to \infty$ $\mathcal{I}_T(V,\beta)$ converges a.s. to the functional $\mathcal{I}_\infty(V,\beta)$ given by

$$\mathcal{I}_\infty(V,\beta) = -d{D_0}^2\beta + \frac{1}{2}\int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \left(|\beta D_0 p + \nabla V(q)|^2 - 2\langle p, \nabla\nabla V(q)p\rangle\right) \varrho_0(q,p)dqdp \tag{3.11}$$

Using the fact that $\varrho_0(q,p)$ is a product of two densities, $\varrho_0(q,p) = \rho_0(q)g_0(p)$, and that $g_0(p)$ is Gaussian, the integral over the momentum in (3.11) can be performed explicitly. The result can be written as

$$\mathcal{I}_\infty(V,\beta) = \frac{1}{2}\int_{\mathbb{R}^d}\Big(|\nabla V(q))|^2 - 2{\beta_0}^{-1}\Delta V(q)\Big)\rho_0(q)dq + d{D_0}^2\left(\tfrac{1}{2}\beta^2/\beta_0 - \beta\right) \tag{3.12}$$

The integral on the right-hand side is, up to an irrelevant constant, the same as the one in (1.15) and it is the only term involving V. As a result, the minimum of (3.12) over V is reached when $V = V_0 + C$, where C is an arbitrary constant. Similarly, the last term in (3.12) is minimized when $\beta = \beta_0$. Thus we conclude that, in the limit as $T \to \infty$, the log likelihood function for V_0 and β_0 has these parameters as unique maximizers.

When T is finite, however, we need to proceed differently. First, we can replace the time integral in (3.10) by an expectation with respect to the occupation measure of the process $\{Q_t, P_t\}_{t\in[0,T]}$:

$$\begin{aligned}\mathcal{I}_T(V,\beta) &= \frac{1}{T}\Big[\beta D_0|P_t|^2 + \langle \nabla V(Q_t), P_t\rangle\Big]_0^T \\ &\quad - d{D_0}^2\beta + \frac{1}{2}\int_{\mathbb{R}^d\times\mathbb{R}^d}\left(|\beta D_0 p + \nabla V(q)|^2 - 2\langle p, \nabla\nabla V(q)p\rangle\right)\mu_T(dq,dp)\end{aligned} \tag{3.13}$$

Assuming that T is large enough so that we can neglect the boundary terms in (3.10) we are left with the terms on the second line in (3.13). To regularize them, we must regularize $\mu_T(dq,dp)$ by some $\varrho_T(q,p)dqdp$. Consistent with (3.4), we assume that the empirical density $\varrho_T(q,p)$ factorizes as $\varrho_T(q,p) = \rho_T(q)g_T(p)$, where $\rho_T(q)$ and $g_T(p)$ are densities which can be estimated separately by splitting the data into $\{Q_t\}_{t\in[0,T]}$ and $\{P_t\}_{t\in[0,T]}$. Consistent with (3.5), we can further assume that $g_T(p)$ is a Gaussian density of the form

$$g_T(p) = (2\pi\beta_T)^{-d/2}e^{-\frac{1}{2}\beta_T|p|^2} \tag{3.14}$$

where $\beta_T > 0$ is a parameter which can be estimated from the data as

$$\beta_T^{-1} = \frac{1}{dT}\int_0^T |P_t|^2 dt. \tag{3.15}$$

Substituting $\varrho_T(q,p)dqdp$ for $\mu_T(dq,dp)$ in the integral term in (3.13) and using (3.14), the integral over the momentum can be performed explicitly. This gives the following approximation for the terms on the second line in (3.13):

$$\frac{1}{2}\int_{\mathbb{R}^d}\Big(|\nabla V(q))|^2 - 2\beta_T^{-1}\Delta V(q)\Big)\rho_T(q)dq + d{D_0}^2\left(\tfrac{1}{2}\beta^2/\beta_T - \beta\right) \tag{3.16}$$

This functional is similar to (3.11), except that it involves the empirical ρ_T and β_T instead of the actual ρ_0 and β_0. The following theorem is thus analogous to Theorem 2.6.

THEOREM 3.2. *Let $U \subset \mathbb{R}^d$ be open and bounded. Suppose that ρ_T is bounded below on U, i.e. $\exists\varepsilon > 0\,\forall x \in U : \rho_T(x) > \varepsilon$ holds. Assume furthermore that $\beta_T > 0$. Then the functional*

$$\tilde{\tilde{I}}_h(V,\beta) = \frac{1}{2}\int_U\left(|\nabla V(q)|^2 - 2\beta_T^{-1}\Delta V(q)\right)\rho_T(q)dq + dD_o^2\left(\frac{1}{2}\beta^2/\beta_T - \beta\right) \tag{3.17}$$

has a unique minizer (V,β) in $\bar{H}^1(U)\times\mathbb{R}$, where the bar denotes functions of mean zero. This minimizer is given by

$$\hat{V} = -\beta_T^{-1}\log\rho_T(x) + C, \qquad \hat{\beta} = \beta_T, \tag{3.18}$$

where the constant C is such as to ensure that $\hat{V}$ has mean zero.

Proof. First establish that $\hat{\beta} = \beta_T$ which is straightforward as β only occurs in the second term. The rest of the proof proceeds analogously to Theorem 2.6. □

Thus, the **non-parametric approach** can be generalized to the Langevin equation and leads to the fitting of V to the empirical measure, similarly to what we found in the case of the Smoluchowski equation. Furthermore, the inverse temperature β is estimated from the variance of the momentum in the empirical measure.

4. Numerical Experiments.

4.1. Setup. In this section we perform a series of numerical experiments on a simple model system to illustrate the results obtained in the previous sections, in particular the relationship between the practitioners' and statisticians' approach to drift estimation. These experiments will also allow us to investigate two issues that we have left open so far. The first is what is the impact in the **parametric approach** of having a data set sampled at discrete points in time rather than continuously? The second issue is how to obtain the approximate density $\rho_T(x)$ needed in the **non-parametric approach**. The model system we will investigate is the one-dimensional diffusion

$$\dot{X}_t = -X_t^3 + \frac{3}{2}X_t + \frac{3}{2}\dot{W}_t, \quad X_0 = 0. \tag{4.1}$$

This equation is a special case of the Smoluchowski equation (1.1) with

$$V_0(x) = \frac{1}{4}x^4 - \frac{3}{4}x^2 \tag{4.2}$$

and $\beta = 8/9$. To generate the data, we integrate (4.1) using the Euler-Maruyama scheme with time-step Δt for $N_T = \lfloor T/\Delta t \rfloor$ steps, i.e. using

$$X_{(j+1)\Delta t} = X_{j\Delta t} - X_{j\Delta t}^3 \Delta t + \frac{3}{2}X_{j\Delta t}\Delta t + \frac{3}{2}\sqrt{\Delta t}\,\xi_j, \quad j = 0, \ldots, N_T - 1, \tag{4.3}$$

with $X_0 = 0$ and where $\{\xi_j\}_{j=0,\ldots,N_T-1}$ are independent Gaussian variables with mean 0 and variance 1. The value of Δt and T will be varied to measure the impact of these parameters. The Euler-Maruyama scheme produces a discrete time sample $\{X_{j\Delta t}\}_{j=0,\ldots,N_T}$ which we will use as data. For simplicity, we will denote this data set as $\{X_j\}_{j=0,\ldots,N_T}$ in the sequel.

In the **parametric approach** we use the following polynomial representation of the force $b_0(x) = -V_0'(x) = -x^3 + \frac{3}{2}x$:

$$b(x, \theta) = \sum_{i=0}^{3} \theta_i x^i. \tag{4.4}$$

Equivalently, this means that we parametrize the potential $V_0(x)$ as

$$V(x, \theta) = \sum_{i=0}^{3} \frac{\theta_i x^{i+1}}{i+1}. \tag{4.5}$$

Based on this parametrization, and consistent with the time-discretization used in (4.3), we adopt the following discretized version of the log likelihood function (1.13)

$$\mathcal{I}_T(\theta) = \frac{1}{T}\sum_{j=0}^{N} \left(|b(X_j, \theta)|^2 \Delta t - 2b(X_j, \theta)\,(X_{j+1} - X_j) \right). \tag{4.6}$$

The minimization of (4.6) gives rise to a linear algebraic system for $\theta = (\theta_0, \ldots, \theta_3)$ which is easy to solve (the solution is similar to (2.23) in the continuously-sampled case). We refer to this solution as the MLE $\hat{\theta}$.

In the **non-parametric approach** the main issue is the evaluation of the empirical density $\rho_T(x)$ in (1.15) and (1.17). To obtain results that can be easily compared with those of the **parametric approach** we will parametrize ρ_T as

$$\rho_T(x,\theta) = Z^{-1}(\theta)e^{-\beta V(x,\theta)} \quad \text{where} \quad Z(\theta) = \int_{\mathbb{R}} e^{-\beta V(x,\theta)}dx \tag{4.7}$$

and $\beta = 8/9$ is given. To then determine $\rho_T(x,\theta)$, we test and compare three different methods. The first method is based on estimating a discretization of the empirical density obtained by a standard histogram method using an even number K of bins centered at $c_k = 8k/K$ for $k = -K/2, \ldots, K/2$. The bins are spaced equidistantly and the small number of samples outside $[-4, 4]$ are discarded. Denoting by $\hat{\rho}_k$ this empirical density, we then obtain $\theta = (\theta_0, \ldots, \theta_3)$ by minimizing

$$\sum_{k=-K/2}^{K/2} \left|\log \hat{\rho}_k + \beta V(c_k,\theta)\right|^2 \tag{4.8}$$

This objective function is the discrete analog of the L^2 norm of the difference between $-\beta V(x,\theta)$ and the log of a (putative) continuous approximation of the empirical density ρ_k. Note that this is a straightforward least squares problem of dimension K, so this is easily solved by standard methods. We refer to optimising (4.8) as the practitioners' method, and call $\hat{\theta}$ optimising (4.8) the PME.

For the second method, note that in one-dimension, the occupation measure μ_T has the scaled local time $L_T(x)/T$ as density, so one can search the minimizer of

$$\int_{\mathbb{R}} \left|\rho_T(x,\theta) - L_T(x)/T\right|^2 dx \tag{4.9}$$

which measures the L^2 distance between $\rho_T(x,\theta)$ and the scaled local time $L_T(x)/T$.

To adapt this to time-discrete observations, it is possible to expand the square and then approximate the local time as

$$L_T = \frac{T}{N_T}\sum_{j=0}^{N_T} \delta_{X_j}.$$

This results in estimation via minimizing the following objective function over θ:

$$\int_{\mathbb{R}} \rho_T^2(x,\theta)dx - \frac{2}{T}\sum_{j=0}^{N_T} \rho_T(X_j,\theta) \tag{4.10}$$

The third method is based on a coarsened version of (4.10) in which we use $\hat{\rho}_k$ to replace (4.10) by

$$\sum_{k=-K}^{K} \rho_T^2(c_k,\theta) - 2\rho_T(c_k,\theta)\hat{\rho}_k \tag{4.11}$$

Minimizing (4.11) is slightly less accurate than minimizing (4.10), but it is computationally less expensive if the number of bins is significantly smaller than the number of data points in the time-series, $K \ll N_T$. The computational cost involved in minimizing (4.10) compells us to use (4.11), but we study its behaviour for several choices

of K, the number of bins in the histogram. To optimise (4.11) we use steepest descent together with a line search strategy and refer to the optimal $\hat{\theta}$ as the minimum distance estimator (MDE).

More generally, using a histogram as a means of summarising the data not only smoothes the empirical density but also makes optimisation easier. In the case of the estimator (4.8), it is even unclear how this estimator could be used with the unsmoothed discrete time empirical density. Various alternative ways of obtaining a smoothed empirical density $\hat{\rho}$ from the discrete time observations X_j are conceivable. Established methods include kernel density estimators and even nonparametric density estimation.

4.2. Connections via Correlation. In order to establish that the link between the MLE (obtained from (4.6)) and the PME (obtained form (4.8) persists for discretely observed data, we wish to study the stochastic dependency between the PME and the MLE understood as random variables.

Having verified that asymptotic unbiasedness and a suitable decay of variance are indeed observed for our implementation of these estimators, we consider that these results are standard at least for the MLE, so that we do not show them here in detail.

Since applied interest resides in the invariant density and the empirical measure, it seems interesting to first compare MLE and density-based estimators at the level of densities. To do this, we perform numerical simulations using $K = 50$ bins for a final time of $T = 100$ (and $\Delta t = 0.01$) and compute the invariant density $\rho(\hat{\theta}, \cdot)$ induced by MLE estimates $\hat{\theta}$ of $\{\theta_i\}_{i=0}^3$. A typical case is shown in Figure 4.1 and repeated experiments computing the bin-wise correlation of deviations from the true invariant density ρ (whose evaluation at c_k we denote by $\rho_k = \rho(c_k)$), namely

$$\alpha = \frac{\sum_{k=-K/2}^{K/2} (\hat{\rho}_k - \rho_k) \cdot \Big(\rho(\hat{\theta}, c_k) - \rho_k\Big)}{\sqrt{\sum_{k=-K/2}^{K/2} (\hat{\rho}_k - \rho_k)^2} \cdot \sqrt{\sum_{k=-K/2}^{K/2} \Big(\rho(\hat{\theta}, c_k) - \rho_k\Big)^2}},$$

show high correlations as visible in the histogram in Figure 4.2. An MDE or PME that now attempts to fit the empirical density $\hat{\rho}$ or its logarithm using some least squares method would hence be expected to yield drift parameter estimates $\hat{\theta}$ whose deviations from θ are correlated with the MLE estimates' deviations.

To investigate whether this is so, it is useful to note the experimental observation that all three estimators display an approximately Gaussian distribution. We use the final time $T = 160$ and the timestep $\Delta t = 0.002$ and MDE and PME each use $K = 50$ bins throughout. We evaluate $N = 1000$ realisations each of MDE, MLE and PME to produce estimates of $\{\theta_3^{(k)}\}_{k=1}^N$ of θ_3. We then standardise these estimates subtracting the mean and dividing by the standard error. Histograms and Quantile-Quantile-Plots of these parameter estimates are given in Figures 4.5, 4.3 and 4.4 respectively. Furthermore, we apply a Kolmogorov-Smirnov test of normality and report the obtained p-values in these Figures. In all three cases, the observed p-value is above $p = 0.88$ so that the observed evidence against normality using the Kolmogorov-Smirnov test statistic is considered very weak. It should be pointed out that for smaller final times, the distribution of parameter estimates does not approximate a Gaussian as closely as this; theorems on (local) asymptotic normality that can be found for the MLE and MDE in continuous time e.g. in [11] only suggest normality for large final times.

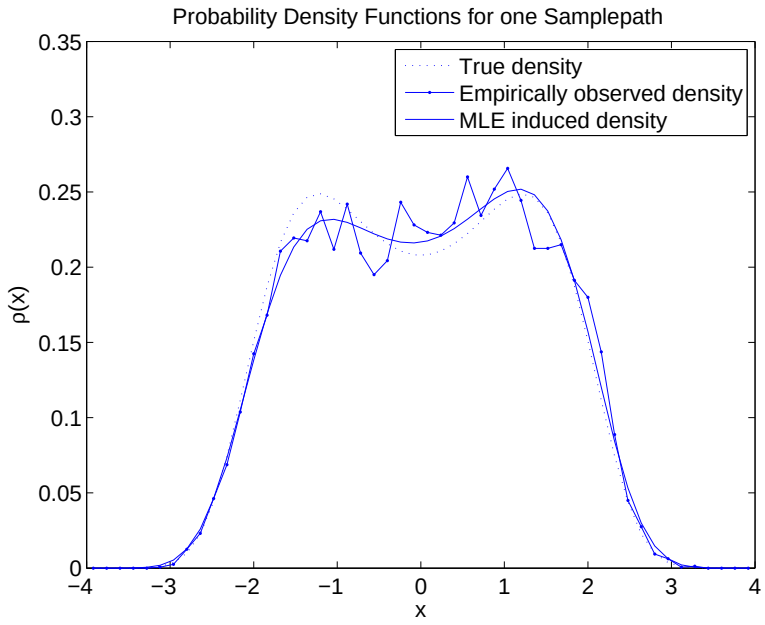

FIG. 4.1. *Densities from one Particular Samplepath*

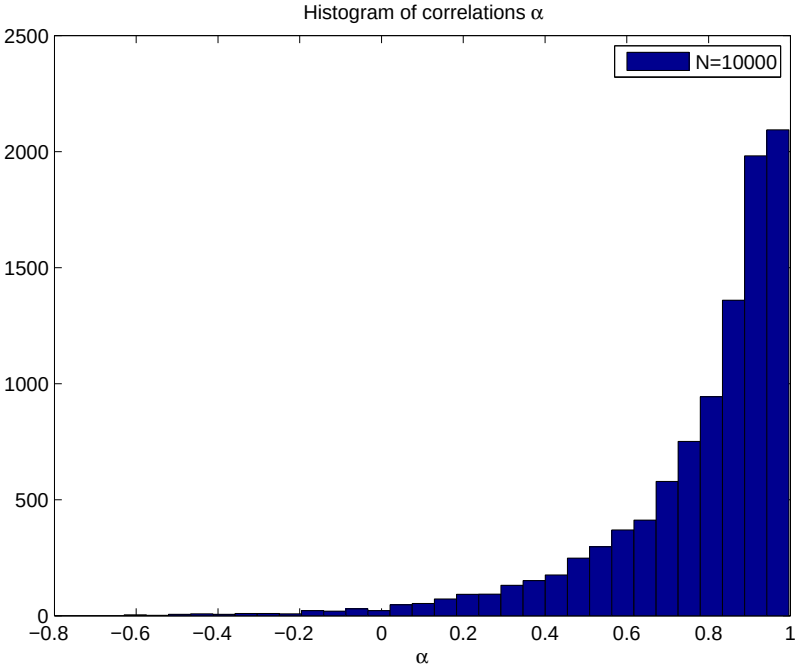

FIG. 4.2. *Correlation coefficients α for deviations of MLE-induced and empirical densities from the invariant density*

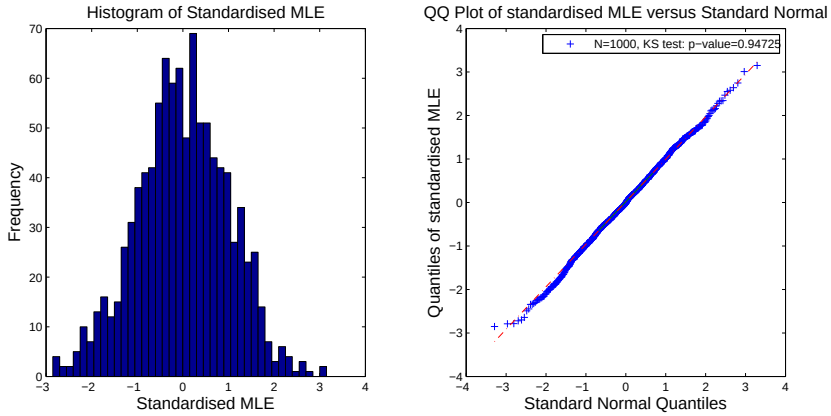

FIG. 4.3. *Test of Normality for the MLE*

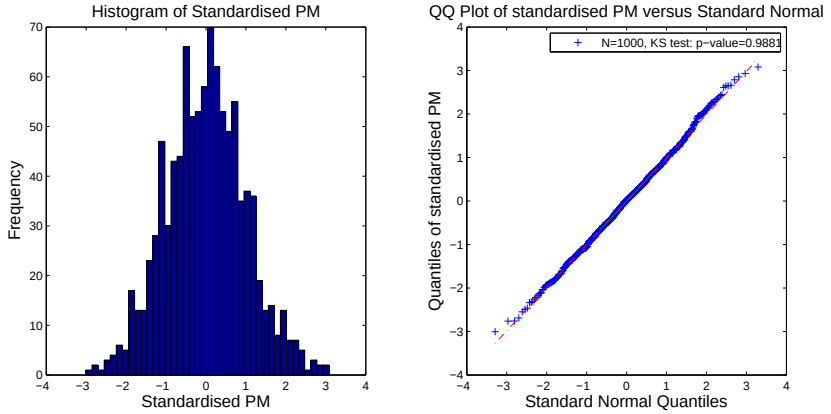

FIG. 4.4. *Test of Normality for the PME*

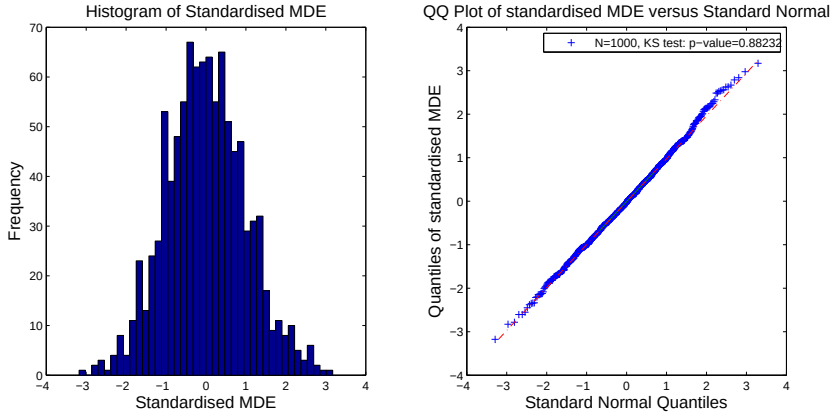

FIG. 4.5. *Test of Normality for the MDE*

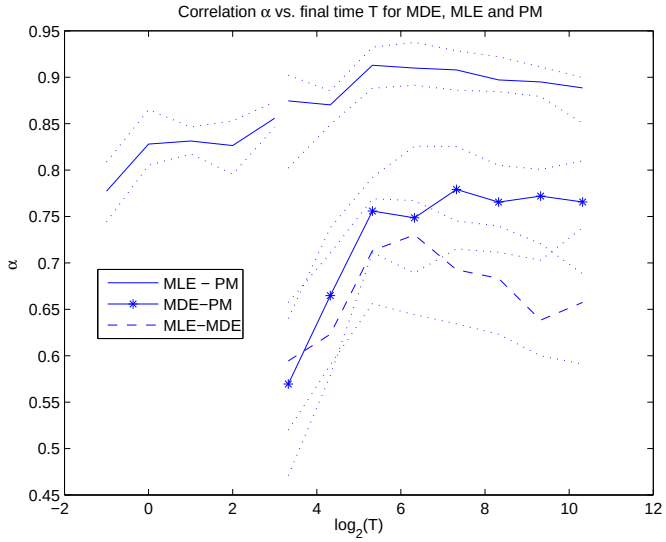

FIG. 4.6. *Correlations of drift parameter deviations for MDE, PME and MLE*
The dotted lines indicate 33% quantile bands.

It is now appropriate to study correlations as a measure of independence, so we consider the deviations of the three estimators of θ_3 from their respective means as a function of final time. Plotting their averaged correlations over at least $N_{av} = 1000$ realisations each as a function of final time T yields the plot in Figure 4.6. It seems that the maximal obtainable correlation coefficient for is around 0.9 for the MLE-PM pair. As would be expected from the analytical link of these estimators, a decline of correlation is observed as the final time T is decreased.

Consulting Figure 4.7, it can be seen that the number of bins has only a small influence on the observed correlation of the correlation between MLE and PME estimates. We view this as an indication that other smoothing methods to arrive at $\hat{\rho}$ would not yield significantly lower correlations.

4.3. Influence of Boundary Conditions at Finite T. The approximation of ignoring boundary terms in going from (1.14) to (1.15) is good in the limit of large final times, as was shown in Theorem 2.5. In this subsection, we will briefly sketch the influence of ignoring these boundary terms for finite, even small final times. To do this most easily, we introduce a variant of the maximum likelihood estimator (abbreviated to MLE2) obtained by minimizing the following objective function:

$$\mathcal{I}_T^{(2)}[\theta] = \sum_{j=0}^{N} \left(|b(X_j, \theta)|^2 \, \Delta t + \sigma^2 b'(X_j, \theta) \Delta t \right). \tag{4.12}$$

Note that this is similar to a discretisation of (2.22) but after having performed a partial integration in the spirit of (2.24) to remove the stochastic integral and neglecting the boundary terms arising from integrating up the resulting Stratonovich integral (whereas the MLE would have been attained by discretising straight away, not performing any partial integrations). It should be compared with $\mathcal{I}_T[\theta]$ in (4.6).

In fact, the deviation of the correlation between MLE2 and MLE from 1 should indicate the influence of the initial-condition (and final value) related term on the

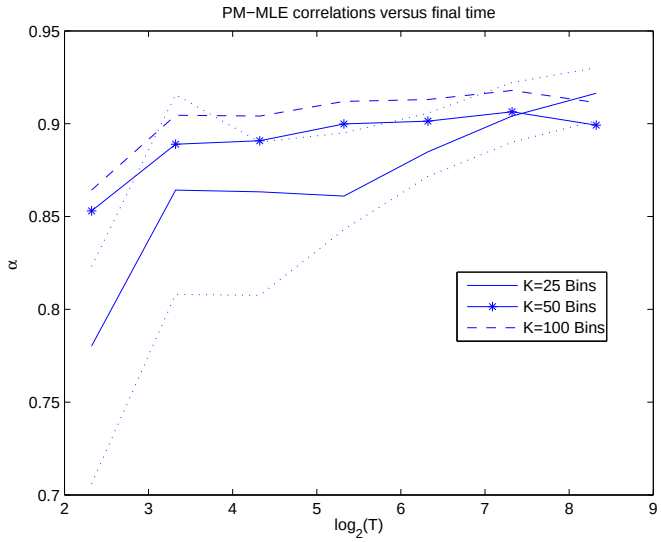

FIG. 4.7. *Correlations of drift parameter deviations for MLE and PME* The dotted lines indicate 33% quantile bands.

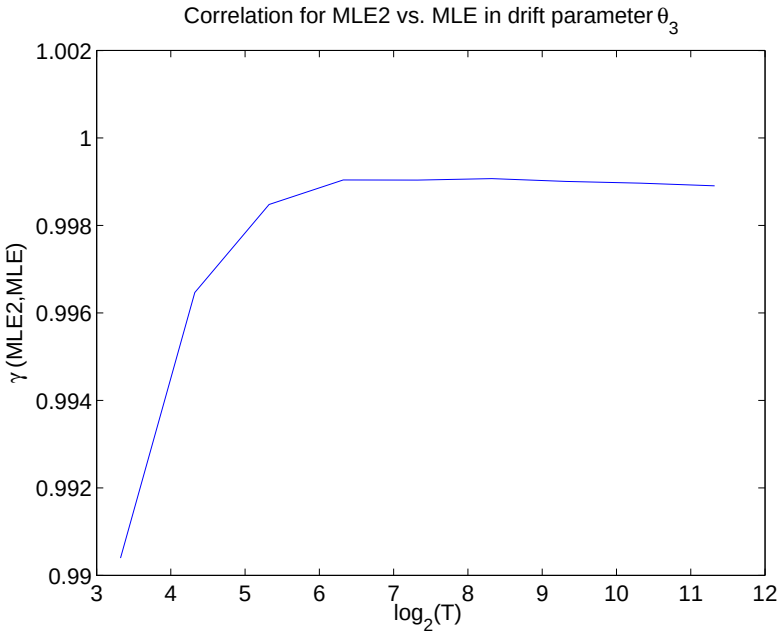

FIG. 4.8. *Correlations of drift parameter deviations for $\tilde{A}$ vs. MLE*

parameter estimates. Using a similar experimental setup (with $\Delta t = 0.0002$ this time), we compute the correlation of the MLE2 estimate and the MLE which results in Figure 4.8.

The remarkably high degree of correlation indicates that the first term which is of order $\mathcal{O}(\frac{1}{T})$ is of little influence for the final times considered in this plot. It does,

however, decline for small final times and the onset of this decline around $T = 10$ is compatible with the decline of correlation observed in Figure 4.6.

5. Conclusions and Future Work. By analyzing different procedures to regularize the likelihood function for the drift of a diffusion, we have highlighted some links between the maximum likelihood principle used widely in the statistical literature, and the practitioners' estimator based on fitting the logarithm of the empirical measure to the drift. These links have been further substantiated through selected numerical examples. In the special case of gradient diffusions these estimators are even more closely linked as their deviations from the mean value satisfy the same statistics to leading order.

At first glance the minimum distance estimator seems to be close to the **non-parametric approach**, but our analysis shows that the link between the **parametric approach** and the **non-parametric approach** is far closer.

This paper leaves open many avenues of further enquiry:

• Our work has been exclusively concerned with reversible problems with equilibrium distribution $e^{-\beta V(q)}$, or non-reversible problems with the equilibrium distribution of the Boltzmann-Gibbs form $e^{-\beta H(q,p)}$, with $H(q,p) = \frac{1}{2}|p|^2 + V(q)$ (separable and quadratic in the momenta). This is natural for examples arising in molecular dynamics. It would also be interesting to perform estimation for processes involving colored noise such as

$$\ddot{Q}_t + \nabla V(Q_t) = B\dot{R}_t$$

where R_t is a suitable m-dimensional Ornstein-Uhlenbeck process involving $\dot{Q}_t$ to satisfy fluctuation dissipation. The process $(Q_t, \dot{Q}_t, R_t)$ then has marginal measure, after integrating out R, of Boltzmann-Gibbs form.

• For problems arising in the e.g. atmospheric sciences [12], more complex distributions will be required. A characterization of the class of stochastic processes for which the link between the **parametric approach** and the **non-parametric approach** can be established would be desirable.

• The option of regularising the likelihood functional (1.11) by including a higher order differential operator to ensure coercivity has been highlighted. This will be pursued for the 1D case in [16] in the framework of Bayesian nonparametric drift estimation.

• Our results rely heavily on the fact that the diffusion coefficient is assumed known. Whilst it is statistical folklore that drift estimation is considerably harder than diffusion estimation (see e.g. [19], [11]), in that the quadratic variation in principle reveals the diffusion coefficient, it is common practical experience with real data that diffusion estimation is the harder problem. This is because the data is often incompatible with a diffusion, or with the desired diffusion, at small time-scales, see e.g. [17]. To overcome this, practitioners often use time-correlation information, or other information concerning $\mathcal{O}(1)$ time-scales, to estimate the diffusion coefficient – see [7], [15] and [22] for example. Furthermore, multiplicative noise models are often appropriate. See [8] and [12], for example, in the context of molecular dynamics and the atmospheric sciences respectively. A systematic nonparametric approach to the problem of diffusion matrix estimation in multiple dimensions and for $\mathcal{O}(1)$ spaced data would be very desirable. See [21] for an overview of parametric diffusion estimation in this context.

[1] F. M. Bandi and P. C. B. Phillips, *Fully nonparametric estimation of scalar diffusion models*, Econometrica, 71 (2003), pp. 241–283.
[2] F. Comte, V. Genon-Catalot and Y. Rozenholc, *Penalized Nonparametric Mean Square Estimation of the Coefficients of Diffusion Processes*, Prpublication MAP5 n2005-21, to appear in Bernoulli.
[3] R. Durrett, *Stochastic Calculus – A practical Introduction*, CRC Press, London (1996).
[4] L. C. Evans, *Partial Differential Equations*, AMS (1998).
[5] C. W. Gardiner, *Handbook of Stochastic Methods*, Springer, Berlin (1985).
[6] E. Gobet, M. Hoffmann and M. Reiss, *Nonparametric estimation of scalar diffusions based on low frequency data*, Ann. Stat., 32 (2004), pp. 2223–2253.
[7] H.Grubmüller, P.Tavan, *Molecular dynamics of conformational substates for a simplified protein model*, J.Chem.Phys., 101 (1994), pp. 5047–5057.
[8] G. Hummer, *Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations*, New Journal of Physics, 7:34 (2005).
[9] J. P. Kahane, *Some random series of functions*, CUP (1985).
[10] O. Kallenberg, *Foundations of Modern Probability*, Springer (1997).
[11] Y. A. Kutoyants, *Statistical Inference for Ergodic Diffusion Processes*, Springer (2004).
[12] A.J. Majda and I. Timofeyev and E. Vanden-Eijnden, *A mathematical framework for stochastic climate models*, Comm. Pure App. Math., 54 (2001), pp. 891–974.
[13] X. Mao, *Stochastic Differential Equations and Applications*, Norwood, second Edition (2007).
[14] J. Mattingly, A. M. Stuart, D. J. Higham, *Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise.* Stoch. Proc. and Applics, 101 (2002), pp. 185-232.
[15] W. Nadler, A. T. Brünger, K. Schulten and M. Karplus, *Molecular and stochastic dynamics of proteins*, Proc. Natl. Acad. Sci., 84 (1987), pp. 7933-7937.
[16] O. Papaspiliopoulos, Y. Pokern, G. O. Roberts, A. M. Stuart, *Bayesian Nonparametric Drift Estimation and Finite Elements*, in preparation, 2007.
[17] G. Pavliotis, A. M. Stuart, *Parameter Estimation for Multiscale Diffusions*, J. Stat. Phys., 127 (2007), pp. 741-781.
[18] N.Privault, A.Réveillac, *Superefficient drift estimation on the Wiener Space*, C. R. Acad. Sci. Paris Ser. I, 343 (2006), pp. 607–612.
[19] B.L.S. Prakasa Rao, *Statistical Inference for Diffusion Type Processes*, Arnold Publishers, London (1999).
[20] H. Risken, *The Fokker Planck Equation*, Springer (1984).
[21] G. O. Roberts, *Exact Simulation and Inference for Diffusions*, Presentation and lecture notes, SemStat (2007).
[22] J. E. Straub, M. Borkovec, B. J. Berne, *Calculation of Dynamic Friction on Intramolecular Degress of Freedom*, J. Phys. Chem., 91:19, (1987), pp. 4995-4998.
[23] D. Williams, *Probability with Martingales*, CUP (1991).

6. Appendix.

Let us consider the random functional

$$\mathcal{I}_B[b] = \int_0^1 b^2(x)w(x) + b'(x)w(x)dx. \tag{6.1}$$

where $b(\cdot) \in H^1(0,1)$ and $w(x)$ is a standard Brownian bridge. We claim that this functional is not bounded below and state this as a theorem:

Theorem 6.1. *There almost surely exists a sequence* $b^{(n)}(\cdot) \in H^1(0,1)$ *such that*

$$\lim_{n\to\infty} \mathcal{I}_B[b^{(n)}] = -\infty \quad \text{a.s.}$$

Proof. For the Brownian bridge we have the representation

$$w(x) = \sum_{i=1}^{\infty} \frac{\sin(i\pi x)}{i}\xi_i \tag{6.2}$$

where the $\{\xi_i\}_{i=1}^{\infty}$ are a sequence of iid normal $\mathcal{N}(0,1)$ random variables. This series converges in $L^2(\Omega; L^2((0,1),\mathbb{R}))$ and almost surely in $C([0,1],\mathbb{R})$, see [9].

Now consider the following sequence of functions $b^{(n)}$:

$$b^{(n)}(x) = \sum_{i=1}^{n} \frac{\xi_i}{i} \cos(i\pi x). \tag{6.3}$$

We think of a fixed realisation $\omega \in \Omega$ of (6.2) for the time being and note that $\{w(x) : x \in [0,1]\}$ is almost surely bounded in $L^\infty((0,1),\mathbb{R})$, so if there exists a $C > 0$ (which may depend on $\{\xi_i\}_{i=0}^{\infty}$) such that

$$\|b^{(n)}\|_{L^2} < C \quad \forall n \in \mathbb{N} \tag{6.4}$$

the first integral in (6.1) will stay finite. By Parseval's identity, it is clear that for the sequence of functionals (6.3) this will be the case if the coefficients $\frac{\xi_i}{i}$ are square-summable.

Computing the second summand in (6.1) is straightforward since the series terminates due to orthogonality:

$$\int_0^1 \left(\sum_{i=1}^{\infty} \frac{\sin(i\pi x)}{i} \xi_i \right) \cdot \left(\sum_{j=1}^{n} \frac{\xi_j}{j} \cos(j\pi x) \right)' dx = -\frac{\pi}{2} \sum_{j=1}^{n} \frac{\xi_j^2}{j}.$$

It can now be seen that (6.1) is unbounded from below if the following two conditions are fulfilled:

$$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j} \xi_j^2 = \infty \tag{6.5}$$

$$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j^2} \xi_j^2 < \infty \tag{6.6}$$

We finally allow ω to vary and seek to establish that the conditions (6.5) and (6.6) are almost surely fulfilled. To do this, first note that the random variables being summed are independent. Thus, by the Kolmogorov 0-1 law the probability for convergence is either zero or one. We proceed by applying Kolmogorov's Three-Series Theorem (theorem 12.5 in [23]) to each of the two sequences to establish (6.5) and (6.6).

We start by treating (6.5). Denote by $X_j\,|^K$ the truncation of the random variable for some $K > 0$ in the sense:

$$X_j\,|^K\,(\omega) = \begin{cases} X_j(\omega) & \text{if } |X_j(\omega)| \le K \\ 0 & \text{if } |X_j(\omega)| > K \end{cases}.$$

To abbreviate notation, define the following two sequences of random variables:

$$X_j = \frac{1}{j} \xi_j^2$$

$$Y_j = \frac{1}{j^2} \xi_j^2$$

Now consider the summability of expected values for the sequence X_j: since ξ_j^2 follows a χ-squared distribution with one degree of freedom, its expected value is one. For

the truncated variable $X_j \mid^K$, for any $K > 0$, there will be some j^* so that for all $j \geq j^*$ we have that

$$\mathbb{E}(X_j \mid^K) = \mathbb{E}\left[\frac{1}{j}\left(\xi^2 \mid^{jK}\right)\right] > \frac{1}{2j}$$

Therefore, the expected value summation fails as follows:

$$\begin{aligned}\sum_{j=1}^{\infty} \mathbb{E}(X_j \mid^K) &= \sum_{j=1}^{\infty} \frac{1}{j} \mathbb{E}\left(\xi^2 \mid^{jK}\right) \\ &\geq \sum_{j=j^*}^{\infty} \frac{1}{2j} = \infty\end{aligned}$$

Therefore, the series $\sum_{j=1}^{\infty} X_j$ diverges to infinity almost surely, thus (6.5) is established.

Now let us establish (6.6) using the Three-series theorem. First check the summability of the expected values:

$$\sum_{j=1}^{\infty} \mathbb{E}(Y_j \mid^K) \leq \sum_{j=1}^{\infty} \mathbb{E} Y_j = \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty$$

Now let us establish the summability of the variances:

$$\begin{aligned}\sum_{j=1}^{\infty} \mathrm{Var}(Y_n \mid^K) &\leq \sum_{j=1}^{\infty} \mathrm{Var} Y_n \\ &= \sum_{j=1}^{\infty} \frac{1}{j^4} \mathrm{Var} \xi_j^2 \\ &= 2 \sum_{j=1}^{\infty} \frac{1}{j^4} < \infty\end{aligned}$$

where we used that ξ_j^2 follows a χ-squared distribution with one degree of freedom and hence has variance $\mathrm{Var}\xi_j^2 = 2$. Finally, to establish the summability of the tail probabilities we use the following argument for any $K > 0$:

$$\begin{aligned}\sum_{j=1}^{\infty} P(|Y_j| > K) &\leq \sum_{j=1}^{\infty} \frac{1}{K} \mathbb{E}|Y_j| \\ &\leq \frac{1}{K} \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty\end{aligned}$$

where we have used the Markov inequality and the previous calculation of the expected value of $Y_j = |Y_j|$.

To put everything together let us reconsider the functional $I[b]$:

$$\begin{aligned} I[b^{(n)}] &= \int_0^1 \left(b^{(n)}\right)^2 (x)w(x) + \left(b^{(n)}\right)' (x)w(x)dx \\ &\leq \left(\sup_{x\in[0,1]} w(x)\right) \int_0^1 \left(b^{(n)}\right)^2 (x)dx - \frac{\pi}{2}\sum_{j=1}^{n} \frac{1}{j}\xi_j^2 \\ &\leq \left(\sup_{x\in[0,1]} w(x)\right) \frac{1}{2}\sum_{j=1}^{n} X_j - \frac{\pi}{2}\sum_{j=1}^{n} Y_j \end{aligned}$$

Now use the almost surely true convergence and divergence statements (6.5) and (6.6) to conclude:

$$\lim_{n\to\infty} I[b^{(n)}] = -\infty \quad \text{a.s.}$$

□